УДК 001.8:002.01

Субботін С. О.

Д-р техн. наук, професор, професор кафедри програмних засобів Запорізького національного технічного університету, Запоріжжя, Україна

МОДЕЛЬ ТА ІНДИВІДУАЛЬНІ МЕТРИКИ ЯКОСТІ НАУКОВИХ ПУБЛІКАЦІЙ

Проаналізовано відомі метрики наукових публікацій. Встановлено, що їх загальним недоліком є оцінювання статей на зовнішньому по відношенню до їхнього вмісту рівню, що не дозволяє явно судити про якість подання матеріалу статті. Метою даної роботи була розробка комплексу показників, які дозволяють характеризувати властивості наукових праць з точки зору їх структури, лексики та форми подання матеріалу, а також бібліографії. Визначено набір метрик, що дозволяють кількісно оцінювати індивідуальні властивості матеріалу і автоматизувати аналіз наукових публікацій. Запропонований набір містить показники статті: потенціал охоплення читацької аудиторії, структурованість, різноманітність географії, мов та видів джерел бібліографії, якість бібліографії, насиченість тексту посиланнями на джерела, число і цитованість рисунків і таблиць, ілюстрованість, обсяг і використання математичного апарату, ефективність абревіатур, показники лексики статті (відповідність назви тексту, назви та авторської анотації, авторської анотації і тексту, ключових слів і анотації, ключових слів і назви, ключових слів і тексту, ключових і характеристичних слів, анотації та характеристичних слів, назви і характеристичних слів, УДК та лексики статті, опису таблиць та рисунків тексту, відповідності лексики абзаців, назв літературних джерел і тексту), самоцитування авторами статті, якість авторського колективу, гібридні та інтегральні показники якості статті. Наведено приклади, що підтверджують практичну застосовність запропонованих показників.

Ключові слова: наукометрія, бібліометрія, якість, стаття, наукова робота, метрика, аналіз цитованості, важливість статті.

НОМЕНКЛАТУРА

IPP – Impact per Publication;

SNIP – Source normalized impact per paper;

SJR – SCImago Journal Rank;

УДК – Універсальна десяткова класифікація;

$\alpha_1(i)$ – коефіцієнт наукового ступеня i-го автора;

$\alpha_2(i)$ – коефіцієнт вченого звання i-го автора;

$\alpha_3(i)$ – коефіцієнт посади i-го автора;

$\alpha_4(i)$ – коефіцієнт рівня організації, де працює i-й автор;

$\alpha_5(i)$ – сукупна кількість посилань на статті i-го автора у наукометричних і реферативних базах;

$\alpha_5(i)$ – середній індекс Гірша i-го автора у наукометричних базах;

a_i^{str} – функція, що повертає кількісну оцінку наявності i-го необхідного елемента у структурі j-ї статті;

abstract – словник авторської анотації j-ї статті;

$abstract^i$ – словник i-ї анотації до j-ї статті;

$abstract_i$ – i-те слово у словнику авторської анотації j-ї статті;

$Area_j$ – обсяг, займаний виданою статтею у см^2;

auditory(x) – функція, що повертає оцінку кількості людей, що володіють мовою x;

aut_k – ідентифікатор k-го автора j-ї статті;

b_1 – рік виходу j-ї статті;

b_2 – кількість робіт авторів j-й статті в у бібліографії до неї (самоцитування);

b_3 – найбільш ранній рік джерела у бібліографії до j-ї статті;

b_4 – найпізніший рік джерела в бібліографії до j-ї статті;

Base – набір наукових праць;

cwrds – набір характеристичних слів j-ї статті;

$cwrds_i$ – i-те слово з набору характеристичних слів j-ї статті;

el_{ji} – формальне позначення i-го елемента j-ї роботи;

eq – функція, що ідентифікує авторство статей;

figcap – словник тексту вмісту k-го рисунка і його підпису;

$figcap_{k,i}$ – i-те слово в словнику тексту вмісту k-го рисунка і його підпису;

geo(ref_i) – функція, що повертає географічний ідентифікатор видавництва i-го посилання у списку бібліографії j-ї статті;

I_j^q – q-й показник якості j-ї статті;

keywords – словник авторських ключових слів j-ї статті;

$keywords_i$ – i-те авторське ключове слово у словнику;

lang(ref_i) – функція, що повертає ідентифікатор мови i-го посилання у списку бібліографії j-ї статті;

lang(x) – функція, що повертає код мови словника x;

maxcit – сума максимальних кількостей цитувань по всіх базах;

N – кількість статей;

N_a – кількість слів у словнику анотації j-ї статті;

N_{abr} – кількість абревіатур (крім загальновідомих) без повторень, використовуваних у j-й статті;

N_{abst} – кількість анотацій у j-й статті;

N_{aut} – кількість авторів j-ї статті;

N_{cwrds} – кількість характеристичних слів j-ї статті;

N_{dmsb} – кількість позначень змінних, констант, нестандартних функцій і операторів, використаних у формулах статті і розшифрованих у тексті статті;

N_{elj} – кількість елементів j-ї статті;

N_f – кількість нумерованих формул у тексті статті;

N_{fig} – кількість рисунків у статті;

N_{figcap_k} – кількість слів у словнику тексту вмісту k-го рисунка і його підпису;

N_{iss} – кількість статей, опублікованих у відповідному номері журналу;

N_{kw} – кількість авторських ключових слів у словнику *keywords*;

N_{msb} – кількість позначень змінних, констант, нестандартних функцій і операторів, використовуваних у формулах статті;

DOI 10.15588/1607-3274-2015-3-4

N_n – кількість слів у словнику назви *j*-ї статті;

N_{par_i} – кількість слів у словнику *i*-го абзацу;

Nre_i – кількість разів, коли *i*-те джерело цитувалося у тексті *j*-ї статті окремо від інших джерел;

$Nrec_i$ – кількість разів, коли *i*-те джерело цитувалося у тексті *j*-ї статті разом з іншими джерелами;

N_{reris_i} – кількість посилань у тексті статті на *i*-й рисунок;

N_{retab_i} – кількість посилань у тексті статті на *i*-ту таблицю;

N_{ref} – кількість джерел у бібліографії до *j*-ї статті;

N_{ris} – кількість рисунків у статті;

N_{rf_i} – кількість посилань на *i*-ту нумеровану формулу в тексті статті;

$N_{sereris_i}$ – кількість посилань у тексті статті на *i*-й рисунок;

$N_{seretab_i}$ – кількість окремих посилань у тексті статті на *i*-ту таблицю;

N_{serf_i} – кількість окремих посилань на *i*-ту нумеровану формулу у тексті статті;

N_t – кількість слів у словнику тексту *j*-ї статті;

N_{tab} – кількість таблиць у статті;

N_{tabcap_k} – кількість слів у словнику тексту вмісту *k*-ї таблиці та її заголовка;

N_{UDC} – кількість слів у словнику УДК;

N_{wref} – кількість слів у списку слів з назв джерел у списку літератури;

$paper_j$ – *j*-та наукова праця;

par_i – словник *i*-го абзацу;

$par_{i,p}$ – *p*-те слово словника *i*-го абзацу;

population – загальна чисельність населення Землі;

Q – набір показників, що характеризують властивості статті;

$reab_i$ – кількість використань *i*-ї абревіатури в *j*-й статті;

$reaut_i$ – список ідентифікаторів авторів *i*-го джерела в бібліографії *j*-ї статті;

ref – список слів з назв джерел у списку літератури *j*-ї статті;

ref_i – *i*-те слово зі списку слів з назв джерел у списку літератури;

sen – кількість речень у тексті *j*-ї статті;

senre – кількість речень у тексті *j*-ї статті, що містить посилання на джерела;

$syn(v_i)$ – функція, що повертає список слів-синонімів для слова v_i, а також їхні переклади (з їхніми синонімами) на всі доступні мови;

t – період часу;

tabcap – словник тексту вмісту *k*-ї таблиці та її заголовка;

$tabcap_{k,i}$ – *i*-те слово в словнику тексту вмісту *k*-ї таблиці та її заголовка;

text – словник тексту статті;

text – словник тексту *j*-ї статті;

time – функція, що повертає рік, том, номер для статті-аргументу;

title – словник назви *j*-ї статті;

$title_i$ – *i*-те слово у словнику *title*;

$type(ref_i)$ – функція, що повертає тип *i*-го джерела бібліографії до *j*-ї статті;

UDC – словник текстових розшифровок наведених у *j*-й статті індексів УДК;

UDC_i – *i*-те слово зі словника текстових розшифровок УДК;

ν – словник;

$ν_i$ – *i*-те слово у словнику *ν*;

w_i^{str} – ваговий коефіцієнт, що характеризує важливість *i*-го елемента структури статті;

$year_i$ – рік *i*-го джерела бібліографії до *j*-ї статті.

ВСТУП

Постійно зростаючий обсяг наукових публікацій робить практично неможливим їхній аналіз людьми вручну і, відповідно, викликає необхідність автоматизації аналізу властивостей наукових публікацій.

Державні структури (у т. ч. фонди, спеціалізовані вчені ради із захисту дисертацій, атестаційні органи), дослідницькі організації, а також приватні фонди, що фінансують дослідження, зацікавлені у використанні математичного забезпечення, що дозволяє аналізувати значимість як окремих публікацій, так і їхніх авторів, організацій авторів, видань, що у єдиній системі критеріїв дає можливість оцінити і порівняти різні дослідження і дослідників, не затрачаючи значних засобів і часу на проведення їхньої спеціальної експертизи людьми.

Для науковців таке математичне забезпечення повинне забезпечувати можливість підтримки прийняття рішень на вибір наукових журналів і конференцій для опублікування своїх досліджень, а також можливість оцінки і порівняння характеристик своїх публікацій і публікацій колег, що стимулює конкуренцію дослідників, їхніх досліджень і дослідних організацій.

Для редакцій наукових журналів також необхідно математичне забезпечення, що дозволяє аналізувати статику і динаміку властивостей опублікованих статей, а також характеризувати якість розглянутих наукових публікацій у єдиній системі показників.

На даний час відомий ряд показників та інструментальних засобів [1–15], що і їх реалізують та дозволяють аналізувати публікації, які вийшли, їхніх авторів, журнали, а також організації за числом посилань на них в інших публікаціях, або за числом завантажень файлів публікацій у мережі Інтернет.

Недоліком відомих засобів [1–15] є те, що вони не дозволяють оцінити характеристики наукових праць до їхньої публікації на стадії розгляду редакцією наукового журналу й обмежуються тільки посиланнями на статті / авторів / організації / журнали, не приділяючи уваги структурі, лексиці і формі подання матеріалу публікацій.

Метою даної статті є розробка комплексу показників, які дозволять кількісно характеризувати властивості наукових праць з погляду на їхню структуру, лексику і форму подання матеріалу, а також враховуватимуть бібліографію статей і посилання на неї в опублікованих джерелах.

1 ПОСТАНОВКА ЗАДАЧІ

Нехай ми маємо набір наукових праць (статей, тез і т.п.) $Base = \{paper_j\}, j = 1, 2, ..., N$. Кожну наукову працю будемо розглядати як сукупність текстових і графічних елементів $paper_j = \{el_{ji}\}, i = 1, 2, ..., N_{elj}$.

Тоді задача оцінювання якості наукової статті полягає у визначенні набору показників $Q = \{I_j^q\}$, що характеризують властивості статті $paper_j$.

Показники якості статті розділимо на абсолютні – ті, що визначені у вихідних одиницях виміру, а також відносні – ті, що співвіднесені зі значеннями відповідних показників інших статей.

На основі кожного абсолютного показника якості статей, опублікованих за визначений період (визначається роком, томом, номером / випуском) у журналі, визначимо, відповідно:

– відносний показник якості номерів (випусків) журналу за період *t*:

$$I'^{q}_{j} = \frac{I^{q}_{j}}{\bar{I}^{q}},$$

де середній показник якості номерів (випусків) журналу за період *t*:

$$\bar{I}^{q} = \frac{1}{N_{iss}} \sum_{j=1}^{N_{iss}} \{(I^{q}_{j}) \mid time(I^{q}_{j}) \in t\}.$$

– нормований показник якості номерів (випусків) журналу за період *t*:

$$\tilde{I}^{qt} = \frac{I^{qt} - \breve{I}^{qt}}{\hat{I}^{qt} - \breve{I}^{qt}},$$

де $\breve{I}^{qt}$ – мінімальний показник якості номерів (випусків) журналу за період *t*:

$$\breve{I}^{qt} = \min_{j=1,2,...,N_{iss}} \{I^{q}_{j} \mid time(I^{q}_{j}) \in t\},$$

$\hat{I}^{t}_{p}$ – максимальний показник якості номерів (випусків) журналу за період *t*:

$$\hat{I}^{qt}_{p} = \max_{j=1,2,...,N_{iss}} \{I^{q}_{j} \mid time(I^{q}_{j}) \in t\}.$$

2 ЛІТЕРАТУРНИЙ ОГЛЯД

Відомі показники [1–15], що характеризують наукові публікації, можна поділити на декілька груп: метрики на рівні статей, метрики на рівні журналів, метрики на рівні авторів.

Метрики на рівні статей характеризують властивості окремих статей. Найбільш широко використовуваними метриками даної групи є: кількість цитувань статті у певній базі публікацій (чим більше стаття цитується, тим вона вважається важливішою), кількість завантажень файлу статті з певної бази (чим більше завантажень статті, тим вона вважається цікавішою і важливішою), PageRank – кількісна величина, що характеризує «важливість» публікації в мережі Інтернет або базі публікацій, яка визначається за вхідним посиланнями з урахуванням того, як сильно вони рекомендують розглянуту публікацію (чим більше є посилань на публікацію, тим вона вважається більш важливою, а її вага визначається з урахуванням ваг посилань, переданих сторінками, що посилаються). Метрики даної групи оперують винятково кількістю посилань або завантажень статей без урахування вмісту, його структури, обсягу, форми і якості подання. Вони дають лише зовнішню оцінку використання матеріалу статті в цілому, не дозволяючи охарактеризувати її властивості і зробити аналіз умісту.

Метрики на рівні журналів [13–15] характеризують набір статей, опублікованих в одному журналі за визначений період. Найбільше широко використовуваними метриками даної групи є: базові метрики (кількість статей, опублікованих у журналі за певний період, кількість цитувань статей, опублікованих у журналі за певний період, середня кількість цитувань статей, опублікованих у журналі за певний період), імпакт-фактор, індекс оперативності, напівжиття цитувань, SNIP, SJR, IPP.

Імпакт-фактор – чисельний показник важливості наукового журналу, що розраховується за базою Thomson Reuters за визначений часовий період як відношення кількості посилань на статті журналу за рік визначення імпакт-фактора до кількості посилань на статті журналу за декілька (частіше два-чотири) попередніх роки (чим вище імпакт-фактор, тим сильніше вплив статей журналу на інші статті за даний період). Переваги імпакт-фактора: велике охоплення наукової літератури базою Thomson Reuters, публічність, простота у розумінні і використанні, як правило, більш жорстка система резензування у журналів з високим імпакт-фактором. Недоліки імпакт-фактора: невизначеність і неоднозначність взаємозв'язку кількості цитувань і якості конкретної статті, висока залежність імпакт-фактора від часового інтервалу, за який він визначається, велика залежність імпакт-фактора від обсягу і частоти статей у конкретній області науки, непрозорість і монополізація процесу розрахунку імпакт-фактора, вплив суб'єктивної думки експертів при відборі видань, індексованих базою, нерівномірне охоплення видань різних країн і на різних мовах [13–15].

Зведений імпакт-фактор (aggregate impact factor) – відношення кількості цитат для всіх журналів у предметній області до числа статей з усіх журналів у предметній області (чим більше його значення, тим більше середня кількість цитат статей у визначеній предметній області) [14, 15].

Індекс оперативності (immediacy index) – кількість цитувань статей журналу за даний рік, поділена на кількість статей, опублікованих у даному році (чим вище значення даного індексу, тим оперативніше цитуються статті, що публікуються) [14, 15].

Напівжиття цитувань (cited half-life) – середній вік статей, що цитувалися щороку, зокрема у Journal Citation Reports, тобто кількість років, що безпосередньо передують року розрахунку індексу з його урахуванням, на які приходиться половина цитованих статей з даного журналу у році розрахунку індексу, а інша половина – цитати статей за попередні до цього періоду роки (чим вище значення даного індексу, тим довше цитуються опубліковані статті) [14, 15].

Вплив публікації (IPP) – кількість цитувань за заданий рік наукових праць, опублікованих за три попередніх роки, поділена на кількість наукових статей, опублікованих за ці ж три роки [1].

SNIP (нормований вплив джерела на статтю) визначається як відношення середньої кількості цитувань статей, опублікованих у журналі, до частки потенціалу цитування і середнього потенціалу цитування статей із предметної області журналу, де потенціал цитування області

визначається як середня довжина списків літератури в області. SNIP вимірює вплив контекстного цитування за допомогою зважування цитувань, заснованого на загальній кількості цитувань у предметній області. Вплив одного цитування одержує більше значення в тих предметних областях, де цитування бувають рідше і навпаки [14, 15].

SJR – це кількість зважених цитувань, отриманих у заданому році, статей, опублікованих у заданому журналі за три попередніх роки, поділене на загальну кількість статей, опублікованих у журналі за три попередніх роки. SJR – міра наукового впливу журналів, що враховує як кількість цитувань, отриманих журналом, так і важливість або престиж журналів, звідкіля отримані цитування [2–5].

Власний фактор (eigenfactor) журналу в заданому році визначається як відсоток зважених цитат, отриманих журналом у заданому році, статей, опублікованих за п'ять попередніх років, від загального числа цитувань, отриманого всіма журналами в базі даних. При цьому не враховуються цитати, отримані журналом, для якого визначається власний фактор [6].

Вплив статті (Article Influence) визначається як частка від ділення Eigenfactor на відсоток усіх статей, зареєстрованих у Journal Citation Reports, що були опубліковані в заданому журналі [6].

Загальним недолікрм метрик даної групи є їхня незастосовність для окремих статей, для неопублікованих робіт, що знаходяться на розгляді, для окремих учених.

Метрики на рівні авторів характеризують вплив авторів на наукові публікації, що наявні у базі. До метрик даної групи відносяться: базові метрики (сума відношень кількості цитувань статей до кількості авторів статей, середня кількість авторів статті, середня кількість цитувань на автора у рік, середня кількість статей на автора), *h*-індекс, *g*-індекс та *i*-індекс й ін. індекси, визначені на їхній основі.

Індекс Гірша (*h*-індекс) – кількісна характеристика продуктивності вченого, групи вчених, наукової організації або країни у цілому, заснована на кількості публікацій і кількості цитувань цих публікацій: учений має індекс *h*, якщо *h* з його *N* статей цитуються як мінімум *h* разів кожна, у той час як (*N*–*h*) статей, що залишилися, цитуються не більш, ніж *h* разів кожна, тобто вчений з індексом *h* опублікував *h* статей, на кожну з яких послалися як мінімум *h* разів. Переваги індексу Гірша: чим вище індекс Гірша, тим більше публікацій написав учений і тем більше на них є посилань. Недоліки індексу Гірша: його залежність від виду, обсягу й охоплення використовуваної бази публікацій, залежність від способу підрахунку (з урахуванням і без урахування самоцитування) [7, 13–15].

Індекс Еггха (*g*-індекс) – кількісний показник для вимірювання наукової продуктивності, що розраховується на основі розподілу цитувань, отриманих публікаціями вченого для заданого набору статей, відсортованого у порядку убування кількості цитувань, що одержали ці статті, тобто *g*-індекс – це найбільшпе число, таке, що *g* самих цитованих статей одержали (сумарно) не менше g^2 цитувань. Колективний індекс g_1 визначають за набором вчених, впорядкованому за убуванням їхніх *g*-індексів, як (унікальне) найбільше число, таке що верхні g_1 дослідників мають у середньому як мінімум *g*-індекс g_1. Перевагою *g*-індексу є те, що він спрямований на поліпшення на *h*-індексу, наділяючи великою вагою високо цитовані статті. Недоліком *g*-індексу є його непристосованість до ситуації, коли середнє кількість цитувань для всіх опублікованих робіт перевищує загальну кількість опублікованих робіт [8–10].

I-індекс – чисельна характеристика публікаційної активності наукової організації, що розраховується на основі розподілу індексу Гірша вчених з даної наукової організації: наукова організація має індекс *i*, якщо не менш *i* учених з цієї організації мають *h*-індекс не менший за *i* [11, 12].

Загальним недоліком метрик даної групи є те, що вони не враховують характеристики вмісту статті, якість його подання.

Узагальнюючи результати аналізу усіх відомих індексів можна дійти висновку, що їхнім загальним недоліком є оцінювання статей на зовнішньому стосовно їхнього вмісту рівні. Це не дозволяє явно судити про якість подання матеріалу статті, а також визначати рівень статті за відсутності доступу до конкретних баз наукових публікацій. Крім того, відомі метрики є сильно залежними від обсягу і ширини охоплення наукових публікацій використовуваними базами, їхньої мови.

Тому представляється необхідним розробити комплекс індивідуальних метрик якості для автоматизації аналізу наукових публікацій, що дозволяють усунути відзначені недоліки відомих метрик.

3 МАТЕРІАЛИ І МЕТОДИ

Під словником текстового фрагменту $\nu=\{\nu_j\}$ будемо розуміти набір слів v_i, що містяться у текстовому фрагменті, поданих у нормалізованій формі (наприклад, іменники і прикметники в однині і називному відмінку, дієслова – у невизначеній формі й однині і т.п.) без повторень.

Також визначимо на основі словників мов $syn(\nu_i)$ – функцію, що повертає список слів-синонімів для слова v_i, а також їхні переклади (з їхніми синонімами) на всі доступні мови.

Ключовим аспектом будь-якої публікації є її доступність для читачів.

Потенціал охоплення читацької аудиторії при необмеженому доступі до статті визначимо за формулою:

$$I_{aud\ j} = \frac{auditory(lang(text)) + \frac{1}{N_t N_{abst}} \sum_{i=1}^{N_{abst}} \{N_{a,i} \cdot auditory(lang(abstract^i)) \,|\, lang(abstract^i) \neq lang(text)\}}{population},$$

де оцінку розміру аудиторії носіїв кожної мови *auditory*(*x*) можна одержати з [16].

Запропонований показник буде приймати значення в діапазоні [0, 1]: чим більше буде його значення, тим стаття буде більш доступною за мовою для потенційно більшої читацької аудиторії.

Відповідно до історично сформованої практики написання наукових статей і її відображенням у вимогах ведучих наукових видавництв світу [17] структура статей повинна містити низку обов'язкових елементів, явну відсутність яких варто розглядати як недолік форми подання результатів досліджень. Кількісно виразити відповідність структури статті вимогам пропонується за допомогою показника якості структурованості статті:

$$I_j^{str} = \frac{\sum_{i=1}^{N_{str}} w_i^{str} a_i^{str}}{\sum_{i=1}^{N_{str}} w_i^{str}},$$

де $0 < w_i^{str} \le 1,\ 0 \le a_i^{str} \le 1.$

У найпростішому випадку функції a_i^{str} можуть повертати значення «1» за наявності в статті *i*-го елемента і значення «0» – за його відсутності. При більш складній із програмістської точки зору реалізації функції a_i^{str} можуть також визначати неявне наведення в статті відомостей, що відносяться до *i*-го необхідного елемента структури, явно не зазначеного у ній, і повертати число в діапазоні від нуля до одиниці.

Пропонується визначати функції a_i^{str} програмним способом, а відповідні їм ваги w_i^{str} для елементів статті – відповідно до табл. 1.

Запропонований показник буде приймати значення в діапазоні [0, 1]: чим більше буде його значення, тим більш якісною за структурою є стаття, тобто тим більше вона відповідає вимогам [17].

Таблиця 1 – Основні елементи структури статті і їхні ваги

Елемент структури статті	w_i^{str}
Індекс УДК чи іншої системи класифікації наукових публікацій	1
Назва статті англійською мовою	0,9
Назва статті іншою мовою	0,1
Анотація англійською мовою	0,9
Анотація іншою мовою	0,1
Ключові слова англійською мовою	0,9
Ключові слова іншою мовою	0,1
Автори й організації англійською мовою	0,9
Автори й організації іншою мовою	0,1
Номенклатура (перелік позначень і скорочень)	0,1
Вступ	1
Постановка задачі (проблеми)	0,9
Мета роботи	0,1
Огляд літератури	1
Експерименти	1
Результати	1
Обговорення	1
Висновки	1
Подяки	0,1
Список літератури мовою оригіналу	0,9
Транслітерований список літератури (якщо наведений тільки один список усі джерела в якому подані латиницею, то вважається, що транслітерований список наявний у статті)	0,1
Матеріали і методи	1
Інші підрозділи (мається на увазі наявність одного і більш підрозділів, з назвами, відсутніми в даній таблиці)	0,8

Важливою інформацією про зв'язок статті з іншими роботами, що дозволяє також оцінити рівень поінформованості авторів про рівень науки у відповідній предметній області, є наведена у ній бібліографія.

Показник розмаїтості географії джерел бібліографії визначимо як:

$$I_j^{bib\,geo} = \frac{1}{N_{ref}} \left| \bigcup_{i=1}^{N_{ref}} geo(ref_i) \right|,\ N_{ref} > 0.$$

Запропонований показник буде приймати значення в діапазоні (0, 1]: його значення буде тим більше, чим різноманітніше місця розташування видавництв джерел, наведених у бібліографії.

Показник розмаїтості мов джерел бібліографії:

$$I_j^{bib\,lang} = \frac{1}{N_{ref}} \left| \bigcup_{i=1}^{N_{ref}} lang(ref_i) \right|,\ N_{ref} > 0.$$

Запропонований показник буде приймати значення в діапазоні (0, 1]: його значення буде тим більше, чим різноманітніше мови джерел, наведених у бібліографії.

Показник розмаїтості видів джерел у бібліографії до *j*-ї статті пропонується визначати за формулою:

$$K_{T\,j} = \frac{1}{8} \sum_{k=1}^{8} \left\{ 1 \left| \left(\sum_{i=1}^{b_1} \{1 \mid type(ref_i) = k\} \right) > 0 \right. \right\},$$

де $type(ref_i)$ бере значення: 1 – стаття у журналі, 2 – тези доповіді на конференції чи стаття (розділ) у книзі, 3 – книга, 4 – документ про інтелектуальну власність (патент на винахід або авторське свідоцтво), 5 – дисертація або автореферат, 6 – електронний ресурс, 7 – стандарт, 8 – інші види джерел.

В ідеалі стаття повинна цитувати усі основні види наукових джерел. У цьому випадку коефіцієнт розмаїтості видів джерел буде приймати значення «1». Чим менше буде значення даного коефіцієнта, тим менше різноманітних видів джерел приведено в бібліографії. У найгіршому випадку (за відсутності бібліографії як такої) коефіцієнт дорівнюватиме нулю.

Показник якості бібліографії *j*-ї статті визначимо за формулою:

$$I_j^{bib} = \left(\frac{\left(N_{ref} - b_2\right) I_j^{bib\,lang} I_j^{bib\,geo} K_{T_j}}{{N_{ref}}^2 \left(b_4 - b_3\right)\left(1 + b_1 - b_4\right)} \right) \times$$

$$\times \left(\sum_{i=1}^{N_{ref}} year_i - N_{ref} b_3 \right) \left(\frac{1 - e^{-\frac{N_{ref}}{2}}}{1 + e^{-\frac{b_2}{2}}} \right).$$

Запропонований показник буде приймати значення в діапазоні [0, 1): його значення буде тим більше, чим ближче до року виходу *j*-ї статті найпізніша публікація у бібліографії, чим більше діапазон охоплення публікацій, чим більш різноманітними є географія видавців, мови і роки виходу публікацій, чим більшою є кількість публікацій у бібліографії, чим більш рівномірним за строками видання є публікації в бібліографії.

Показник насиченості тексту статті посиланнями на джерела задамо як:

$$I_j^{cit} = \frac{senre}{sen}\left(\frac{\sum_{i=1}^{N_{ref}} N_{re\,i}}{\sum_{i=1}^{N_{ref}}\left(N_{re\,i}+N_{rec\,i}\right)}\right),\ sen>0.$$

Показник якості цитування буде приймати значення у діапазоні [0, 1]: його значення буде тим більше, чим більшою буде в аналізованій статті частка речень, що містять посилання на джерела, а також чим більшою буде частка окремих посилань на джерела у загальній кількості посилань у тексті.

На основі запропонованих показників якості бібліографії і насиченості тексту статті посиланнями на джерела визначимо гібридний показник якості бібліографії і її використань у тексті статті:

$$I_j^{bc} = I_j^{bib} I_j^{cit}.$$

Запропонований показник буде приймати значення в діапазоні [0, 1): його значення буде тим більше, чим ближче до року виходу *j*-ї статті найпізніша публікація у бібліографії, чим більшим є діапазон охоплення публікацій, чим більш різноманітними є географія видавців, мови і роки виходу публікацій, чим більшою є кількість публікацій у бібліографії, чим більш рівномірним за строками видання є публікації в бібліографії, чим більшою буде в аналізованій статті частка речень, що містять посилання на джерела, а також чим більшою буде частка окремих посилань на джерела у загальній кількості посилань у тексті.

Поряд з бібліографією про рівень подання матеріалу статті може свідчити її ілюстративний апарат, що містить таблиці і рисунки.

Показник кількості і цитовності рисунків визначимо за формулою:

$$I_j^{ris} = \begin{cases} \left(\frac{1-e^{-N_{ris}}}{1+e^{-N_{ris}}}\right)\left(\frac{1}{N_{ris}}\sum_{i=1}^{N_{ris}}\{1 \mid N_{reris\,i}>0\}\right)\times \\ \times\left(\frac{1}{N_{ris}}\sum_{i=1}^{N_{ris}}\left(\frac{N_{sereris\,i}}{N_{reris\,i}}\middle| N_{reris\,i}>0\right)\right), N_{ris}>0, \\ 0, N_{ris}=0. \end{cases}$$

Запропонований показник буде приймати значення в діапазоні [0, 1): його значення буде тим більше, чим більше рисунків є у статті, чим більшою є частка рисунків, на які у тексті статті є посилання, а також чим більше посилань у тексті статті наведено на кожен рисунок окремо.

Показник обсягу і цитовності таблиць:

$$I_j^{tab} = \begin{cases} \left(\frac{1-e^{-N_{tab}}}{1+e^{-N_{tab}}}\right)\left(\frac{1}{N_{tab}}\sum_{i=1}^{N_{tab}}\{1 \mid N_{retab\,i}>0\}\right)\times \\ \times\left(\frac{1}{N_{tab}}\sum_{i=1}^{N_{tab}}\left(\frac{N_{seretab\,i}}{N_{retab\,i}}\middle| N_{retab\,i}>0\right)\right), N_{tab}>0, \\ 0, N_{tab}=0. \end{cases}$$

Запропонований показник буде приймати значення в діапазоні [0, 1): його значення буде тим більше, чим більше таблиць є у статті, чим більшою є частка таблиць, на які в тексті статті є посилання, а також чим більше посилань у тексті статті наведено на кожну таблицю окремо.

Показник ілюстрованості статті пропонується визначати за формулою:

$$I_j^{il} = 0{,}5(I_j^{ris} + I_j^{tab}).$$

Запропонований показник буде приймати значення в діапазоні [0, 1): його значення буде тим більше, чим більше ілюстрацій (рисунків і таблиць) є у статті, чим більшою є частка ілюстрацій, на які в тексті статті є посилання, а також чим більше посилань у тексті статті наведено на кожну ілюстрацію окремо.

За умови реалізації програмної оцінки фактичної площі у см^2, займаної ілюстраціями у статті, $Area_j^{il}$ доцільно використовувати показник, що визначається формулою:

$$I_j^{ilvol} = \frac{Area_j^{il}}{Area_j},\ Area_j>0.$$

Запропонований показник буде приймати значення в діапазоні [0, 1]: його значення буде тим більше, чим більше частка площі, займаної ілюстраціями, у площі статті.

На основі запропонованих показників визначимо гібридний показник ілюстрованості статті:

$$I_j^{ilh} = I_j^{il} I_j^{ilvol}.$$

Важливою складовою будь-якої наукової статті є її нотація і математичний апарат.

Показник обсягу і використання математичного апарату статті пропонується визначати за формулою:

$$I_j^{eq} = \begin{cases} \left(\frac{1-e^{-N_f}}{1+e^{-N_f}}\right)\left(\frac{1}{N_f}\sum_{i=1}^{N_f}\{1 \mid N_{rf\,i}>0\}\right)\times \\ \times\left(\frac{1}{N_f}\sum_{i=1}^{N_f}\left(\frac{N_{serf\,i}}{N_{rf\,i}}\middle| N_{rf\,i}>0\right)\right)\left(\frac{N_{dmsb}}{N_{msb}}\right), N_f>0, \\ 0, N_f=0. \end{cases}$$

Запропонований показник буде приймати значення в діапазоні [0, 1): його значення буде тим більше, чим більше в тексті статті нумерованих формул, чим на більше число нумерованих формул є посилань у тексті статті, чим більше частка окремих посилань на кожну формулу у загальній кількості посилань на формули, а також чим більше використаних нестандартних позначень розшифровано у тексті статті.

Значну інформацію про рівень і зміст наукової статті несе її лексика.

Показник ефективності введених абревіатур у *j*-й статті визначимо за формулою:

$$I_{abb\,j} = \begin{cases} \dfrac{1}{N_{reabr}} \sum\limits_{i=1}^{N_{abr}} \{reab_i \mid reab_i > 2\}, N_{reabr} > 0; \\ 0, N_{reabr} = 0, \end{cases}$$

$$N_{reabr} = \sum_{i=1}^{N_{abr}} reab_i.$$

Даний показник буде приймати значення в діапазоні [0, 1]: чим більше буде його значення, тим більш обгрунтованим є використання введених абревіатур у *j*-й статті.

Відповідність назви тексту статті визначимо за допомогою показника:

$$I_{nt_j} = \frac{1}{N_n} \sum_{i=1}^{N_n} \{1 \mid syn(title_i) \cap text \neq \varnothing\}.$$

Цей показник буде приймати значення в діапазоні [0, 1]: чим більше буде його значення, тим більше лексика назви відповідає тексту статті.

Відповідність назви статті авторської анотації статті визначимо за допомогою показника:

$$I_{na_j} = \frac{1}{N_n} \sum_{i=1}^{N_n} \{1 \mid syn(title_i) \cap abstract \neq \varnothing\}.$$

Даний показник буде приймати значення в діапазоні [0, 1]: чим більше буде його значення, тим більше лексика назви відповідає лексиці анотації статті.

Відповідність авторської анотації тексту статті визначимо за допомогою показника:

$$I_{at_j} = \frac{1}{N_a} \sum_{i=1}^{N_a} \{1 \mid syn(abstract_i) \cap text \neq \varnothing\}.$$

Цей показник буде приймати значення в діапазоні [0, 1]: чим більше буде його значення, тим більше лексика авторської анотації відповідає тексту статті.

Відповідність авторських ключових слів і анотації *j*-ї статті можливо оцінити за допомогою показника:

$$I_{ka_j} = \frac{1}{N_{kw}} \sum_{i=1}^{N_{kw}} \{1 \mid syn(keywords_i) \cap abstract \neq \varnothing\}.$$

Даний показник буде приймати значення в діапазоні [0, 1]: чим більше буде його значення, тим більше виділених автором ключових слів або їхніх синонімів чи перекладів зустрічається в тексті авторської анотації до статті.

Відповідність авторських ключових слів і назви *j*-ї статті можливо оцінити за допомогою показника:

$$I_{kn_j} = \frac{1}{N_{kw}} \sum_{i=1}^{N_{kw}} \{1 \mid syn(keywords_i) \cap title \neq \varnothing\}.$$

Даний показник буде приймати значення в діапазоні [0, 1]: чим більше буде його значення, тим більше виділених автором ключових слів або їхніх синонімів чи перекладів зустрічається в тексті назви статті.

Відповідність авторських ключових слів і тексту *j*-ї статті можливо оцінити за допомогою показника:

$$I_{kt_j} = \frac{1}{N_{kw}} \sum_{i=1}^{N_{kw}} \{1 \mid syn(keywords_i) \cap text \neq \varnothing\}.$$

Даний показник буде приймати значення в діапазоні [0, 1]: чим більше буде його значення, тим більше виділених автором ключових слів або їхніх синонімів чи перекладів зустрічається в тексті статті.

Автори не завжди вдало обирають ключові слова. Частково це можна пояснити халатним відношенням або нерозумінням призначення ключових слів слугувати для пошуку статті у базах публікацій.

Визначимо характеристичне слово як таке слово або його синонім чи переклад, що є одним з найбільш частих у статті, або таке слово, що дозволяє відокремити дану статтю від подібних їй за тематикою. Характеристичні слова можуть виділятися без участі людини за допомогою частотного аналізу лексики статей, а також кластеризації бази статей. Характеристичні слова можуть частково збігатися з авторськими ключовими словами і, чим більше таких збігів, тим краще автор виділив ключові слова.

Відповідність виділених автором ключових слів, характеристичним словам статті пропонується оцінювати за формулою:

$$I_{kc_j} = \frac{1}{N_{kw}} \sum_{i=1}^{N_{kw}} \{1 \mid syn(keyworsd_i) \cap cwrds \neq \varnothing\}.$$

Даний показник буде приймати значення в діапазоні [0, 1]: чим більше буде його значення, тим більше авторських ключових слів або їхніх синонімів чи перекладів зустрічається серед характеристичних слів статті.

Аналогічно, відповідність авторської анотації характеристичним словам статті пропонується оцінювати за формулою:

$$I_{ac} = \frac{1}{N_{cwrds}} \sum_{i=1}^{N_a} \{1 \mid syn(abstract_i) \cap cwrds \neq \varnothing\}.$$

Даний показник буде приймати значення в діапазоні [0, 1]: чим більше буде його значення, тим більше характеристичних слів статті або їхніх синонімів чи перекладів зустрічається в тексті авторської анотації до статті, тобто тим можливо вдаліше обрана лексика анотації автором.

Відповідність назви і характеристичних слів *j*-ї статті пропонується оцінювати за допомогою показника:

$$I_{nc_j} = \frac{1}{N_n} \sum_{i=1}^{N_n} \{1 \mid syn(title_i) \cap cwrds \neq \varnothing\}.$$

Даний показник буде приймати значення в діапазоні [0, 1]: чим більше буде його значення, тим більше слів назви або їхніх синонімів чи перекладів зустрічається серед характеристичних слів статті.

Відповідність УДК лексиці статті пропонується оцінювати за допомогою показника:

$$I_{UDC\,j} = \frac{1}{N_{UDC}} \sum_{i=1}^{N_{UDC}} \{1 \mid syn(UDC_i) \cap$$

$$\cap (title \cup abstract \cup keywords \cup text) \neq \varnothing\},$$

словник *UDC* може бути складений на основі [18].

Даний показник буде приймати значення в діапазоні [0, 1]: чим більше буде його значення, тим більше слів з

розшифровок індексів УДК або їхніх синонімів чи перекладів зустрічається серед слів статті.

Відповідність опису таблиць тексту статті пропонується визначати за формулою:

$$I_{tab\,j} = \frac{1}{N_{tab}} \sum_{k=1}^{N_{tab}} \left(\frac{1}{N_{tabcap_k}} \sum_{i=1}^{N_{tabcap_k}} \{1 \mid syn(tabcap_{k,i}) \cap text \neq \varnothing\} \right).$$

Даний показник буде приймати значення в діапазоні [0, 1]: чим більше буде його значення, тим більше слів з тексту і заголовків таблиць зустрічається серед слів статті.

Відповідність опису рисунків тексту статті пропонується визначати за формулою:

$$I_{fig\,j} = \frac{1}{N_{fig}} \sum_{k=1}^{N_{fig}} \left(\frac{1}{N_{figcap_k}} \sum_{i=1}^{Nfigcap_k} \{1 \mid syn(figcap_{k,i}) \cap text \neq \varnothing\} \right).$$

Даний показник буде приймати значення в діапазоні [0, 1]: чим більше буде його значення, тим більше слів з тексту і підписів до рисунків зустрічається серед слів статті.

Показник відповідності лексики *i*-го і *k*-го абзаців *j*-ї статті визначимо як:

$$I_{par\,j}(i,k) = \frac{1}{N_{par_i}} \sum_{p=1}^{N_{par_i}} \{1 \mid syn(par_{i,p}) \cap par_k \neq \varnothing\}.$$

Даний показник буде приймати значення в діапазоні [0, 1]: чим більше буде його значення, тим більш однаковою є лексика відповідних абзаців.

Узагальнений показник відповідності лексики назв літературних джерел лексиці тексту статті визначимо за формулою:

$$I_{rt\,j} = \frac{1}{N_{wref}} \sum_{i=1}^{N_{wref}} \{1 \mid syn(ref_i) \cap text \neq \varnothing\}.$$

Даний показник буде приймати значення в діапазоні [0, 1]: чим більше буде його значення, тим більше слів з назв бібліографічних джерел, їхніх синонімів і перекладів зустрічається серед слів статті.

Показник самоцитування авторами статті визначимо за формулою:

$$I_{sc\,j} = \frac{1}{N_{ref}} \sum_{i=1}^{N_{ref}} \left\{ 1 \left| \left(\sum_{k=1}^{N_{aut}} \{1 \mid eq(aut_k, reaut_i)\} \right) > 0 \right. \right\},$$

де функція, що ідентифікує авторство статей, *eq* повертає значення «1», якщо обидва аргументи ідентифікують одного й того ж автора, «0» – у іншому випадку.

Показник якості авторського колективу статті визначимо як:

$$I_{Qaut\,j} = \left(\frac{1 - e^{-2z}}{1 + e^{-2z}} \right),$$

$$z = \frac{1}{6} \sum_{i=1}^{N_{aut}} \left((\alpha_1(i) + \alpha_2(i) + \alpha_3(i)\alpha_4(i)) \left(\frac{\alpha_5(i) + \alpha_6(i)^3}{maxcit} \right) \right),$$

де $\alpha_1(i)$: 1 – доктор наук (хабілітований), 0,5 – кандидат наук / доктор філософії, 0,1 – магістр наук, інженер, 0 – інше; $\alpha_2(i)$: 1 – професор, 0,5 – доцент / асоційований професор, 0 – інше; $\alpha_3(i)$: 1 – президент, віце-президент, ректор, проректор, декан, заст. декана, завідувач кафедри, начальник відділу, професор, 0,5 – доцент, науковий співробітник, 0,1 – асистент, інженер, фахівець, 0 – інше; $\alpha_4(i)$: 1 – університет, інститут, 0,5 – коледж, технікум, 0,1 – індустрія, 0 – інше.

Інтегральний показник якості наукової статті визначимо на основі показників, що характеризують окремі її властивості:

$$I_{Q\,j} = (I_{aud\,j} I_{Qaut\,j} I_j^{str}) 0{,}5 (I_j^{bc} + (1 - I_{sc\,j})) \times$$
$$0{,}2 (I_j^{il} + I_j^{eq} + I_{abb\,j} + I_{tab\,j} + I_{fig\,j}) \times$$
$$\times 0{,}09 (I_{nt_j} + I_{na_j} + I_{at_j} + I_{ka_j} + I_{kn_j} +$$
$$+ I_{kt_j} + I_{kc_j} + I_{ac} + I_{nc_j} + I_{UDC\,j} + I_{rt\,j}).$$

Даний показник буде приймати значення в діапазоні [0, 1]: чим більше буде його значення, тим вищою буде якість подання матеріалу і лексичний взаємозв'язок елементів статті.

4 ЕКСПЕРИМЕНТИ

Для експериментального дослідження запропонованих показників була розроблена комп'ютерна програма, що автоматично визначає розроблені показники для статті, поданої у вигляді окремого файлу на диску.

Для розрахунків використовувався матеріал даної статті. У табл. 2 наведені деякі вихідні характеристики статті.

5 РЕЗУЛЬТАТИ

Для даної статті за допомогою розробленого програмного забезпечення визначені значення показників, запропонованих у даній роботі, що наведені у табл. 3.

Як видно з табл. 3, запропоновані показники дозволяють характеризувати властивості наукових праць з погляду на їхню структуру, лексику і форму подання матеріалу, а також бібліографію.

6 ОБГОВОРЕННЯ

У порівнянні з відомими показниками, що відбивають лише властивості статей на зовнішньому до їхнього матеріалу рівні [1–15], запропоновані показники дозволяють кількісно оцінювати індивідуальні властивості матеріалу, що дозволяє явно судити про якість подання матеріалу, а також автоматизувати аналіз наукових публікацій.

ВИСНОВКИ

У роботі вирішено завдання автоматизації аналізу якості наукових публікацій за допомогою розробки математичного забезпечення для чисельного оцінювання якості наукових публікацій.

Наукова новизна отриманих результатів полягає у тому, що вперше запропонована модель якості наукових публікацій, що являє собою набір метрик, які дозволяють кількісно оцінювати індивідуальні властивості матеріалу й автоматизувати аналіз наукових публікацій. Запропонований набір містить такі показники статті: потенціал охоплення читацької аудиторії, структурованість, розмаїтість географії, мов і видів джерел бібліографії, якість бібліографії, насиченість тексту посиланнями на джерела, кількість і цитованість рисунків і таблиць, ілюстрованість, обсяг і використання математичного апарату, ефективність абревіатур, показники лексики статті (відповідність назви тексту, назви – авторській анотації, авторській анотації – тексту, ключових слів – анотації, ключових слів – назві, ключових слів – тексту, ключових і характеристичних слів, анотації і характеристичних слів, назви і характе-

Таблиця 2 – Вихідні характеристики статті

Показник	Значення
b_1	2015
b_2	0
b_3	2005
b_4	2014
N_{ref}	18
$\alpha_1(i)$	1
$\alpha_2(i)$	1
$\alpha_3(i)$	1
$\alpha_4(i)$	1
$\alpha_5(i)$	558
$\alpha_6(i)$	4
b_2	0
b_3	2005
b_4	2014
maxcit	30000
N_a	10
N_{abr}	4
N_{abst}	3
N_{aut}	1
N_{ref}	18
N_{fig}	0
N_{tab}	3
N_f	0
N_{reabr}	11
N_f	0
N_{ref}	18
N_t	587
N_{tab}	3
population	$7 \cdot 10^9$
N_{kw}	10
N_n	6
N_{UDC}	20
N_{wref}	93
sen	212
senre	20

Таблиця 3 – Показники якості статті

Показник	Значення
$I_{aud\,j}$	0,0193
I_j^{str}	0,98667
$I_j^{bib\,geo}$	0,5
$I_j^{bib\,lang}$	0,111111
$K_{T\,j}$	0,375
I_j^{bib}	0,0027
I_j^{cit}	0,008359
I_j^{bc}	$2{,}2572 \cdot 10^{-5}$
I_j^{ris}	0
I_j^{tab}	0,90515
I_j^{il}	0,45257
I_j^{eq}	0
$I_{abb\,j}$	0,2727
I_{nt_j}	0,8333
I_{na_j}	0,8333
I_{at_j}	0,8279
I_{ac}	0,7895
I_{ka_j}	0,7
I_{kn_j}	0,4
I_{nc_j}	0,5714
I_{kt_j}	0,8
I_{kc_j}	0,5
$I_{UDC\,j}$	0,4737
$I_{tab\,j}$	0,9
$I_{fig\,j}$	0
$I_{rt\,j}$	0,4625
$I_{sc\,j}$	0
$I_{Qaut\,j}$	0,01037
$I_{Q\,j}$	$2{,}08398 \cdot 10^{-5}$

ристичних слів, УДК і лексики статті, опису таблиць і рисунків – тексту, відповідності лексики абзаців, назв літературних джерел – тексту), самоцитування авторами статті, якість авторського колективу, гібридні й інтегральні показники якості статті. Розроблені показники дозволяють характеризувати властивості наукових праць з погляду на їхню структуру, лексику і форму подання матеріалу, а також бібліографію.

Практична цінність отриманих результатів полягає у розробленому програмному забезпеченні для автоматизації розрахунку запропонованих метрик.

Перспективи подальших досліджень полягають у тому, щоб інтегрувати розрахунок запропонованих показників до наукометричних баз, досліджувати можливі взаємозв'язки запропонованих і відомих бібліометричних показників на великому масиві публікацій, вивчити можливий взаємозв'язок запропонованих показників і експертних оцінок якості статей рецензентами-людьми, що є провідними фахівцями у відповідних областях науки.

ПОДЯКИ

Робота виконана у межах держбюджетної науково-дослідної теми Запорізького національного технічного університету «Інтелектуальні інформаційні технології автоматизації проектування, моделювання, керування і діагностування виробничих процесів і систем» (номер держ. реєстрації 0112U005350) при частковій підтримці міжнародного проекту «Центри передового досвіду для молодих учених» програми «Темпус» Європейської Комісії (№ 544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES).

СПИСОК ЛІТЕРАТУРИ

1. Moed H. F. Measuring contextual citation impact of scientific journals / H. F. Moed // Journal of Informetrics. – 2010. – Vol. 4, Issue 3. – P. 265–277. DOI: 10.1016/j.joi.2010.01.002
2. González-Pereira B. A new approach to the metric of journals' scientific prestige: The SJR indicator / B González-Pereira, V. P. Guerrero-Bote, F. Moya-Anegyn // Journal of informetrics. – 2010. – Vol. 4, Issue 3. – P. 379–391. DOI: 10.1016/j.joi.2010.03.002
3. Guerrero-Bote V. P. A further step forward in measuring journals' scientific prestige: The SJR2 indicator / V. P. Guerrero-Bote, F. Moya-Anegyn // Journal of Informetrics. – 2012. – Vol. 6, Issue 4. – P. 674–688. DOI: 10.1016/j.joi.2012.07.001
4. Butler D. Free journal-ranking tool enters citation market / D. Butler // Nature. – 2008. – Vol. 451, Issue 6. – P. 6. DOI:10.1038/451006a.
5. Falagas M. E. Comparison of SCImago journal rank indicator with journal impact factor // M. E. Falagas, V. D. Kouranos, R. Arencibia-Jorge, D. E. Karageorgopoulos // The FASEB Journal. – 2008. – Vol. 22, Isuue 8. – P. 2623–2628. DOI:10.1096/fj.08-107938.
6. Bergstrom C. T. The Eigenfactor Metrics / C. T. Bergstrom, J. D. West, M. A. Wiseman // Journal of Neuroscience. – 2008. – Vol. 28, Issue 45. – P. 11433–11434.
7. Hirsch J. E. An index to quantify an individual's scientific research output // Proceedings of The National Academy of Sciences. – 2005. – Vol. 102, № 46. – P. 16569–16572. DOI:10.1073/pnas.0507655102.
8. Egghe L. Theory and practise of the g-index / L. Egghe // Scientometrics. – 2006. – Vol. 69, № 1. – P. 131–152. DOI:10.1007/s11192-006-0144-7.
9. Woeginger G. J. An axiomatic analysis of Egghe's g-index / G. J. Woeginger // Journal of Informetrics. – 2008. – Vol. 2, Issue 4. – P. 364–368. DOI:10.1016/j.joi.2008.05.002
10. Tol R. S. J. A rational, successive g-index applied to economics departments in Ireland / R. S. J. Tol // Journal of Informetrics. – 2008. – Vol. 2, Issue 2. – P. 149–155. DOI:10.1016/j.joi.2008.01.001.
11. Kosmulski M. I – a bibliometric index / M. Kosmulski // Forum Akademickie. – 2006. – Vol. 11. – P. 31.
12. Prathap G. Hirsch-type indices for ranking institutions' scientific research output // Current Science. – 2006. – Vol 91, Issue 11. – P. 1439.
13. Игра в цыфирь, или как теперь оценивают труд ученого (сборник статей о библиометрике). – М. : МЦНМО, 2011. – 72 с.
14. Цыганов А. В. Краткое описание наукометрических показателей, основанных на цитируемости // Управление большими системами : сб. тр. Спец. вып. 44 – Наукометрия и экспертиза в управлении наукой / [под ред. Д. А. Новикова, А. И. Орлова, П. Ю. Чеботарева]. – М. : ИПУ РАН, 2013. – С. 248–261.
15. Штовба С. Д. Обзор наукометрических показателей для оценки публикационной деятельности ученого / С. Д. Штовба, Е. В. Штовба // Управление большими системами : сб. тр. Спец. Вып. 44 – Наукометрия и экспертиза в управлении наукой / [под ред. Д. А. Новикова, А. И. Орлова, П. Ю. Чеботарева]. – М. : ИПУ РАН, 2013. – С. 262–278.
16. List of languages by total number of speakers [Electronic resource]. – Access mode: https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers.

17. Рекомендации EASE (Европейской ассоциации научных редакторов) для авторов и переводчиков научных статей, которые должны быть опубликованы на английском языке [Электронный ресурс]. – 16 с. – Режим доступа: http://www.ease.org.uk/sites/default/files/ease_guidelines-june2014-russian.pdf
18. The UDC Summary [Electronic resource]. – Access mode: http://www.udcc.org/udcsummary/php/index.php?lang=ru&pr=Y

Стаття надійшла до редакції 04.02.2015.

Субботин С. А.
Д-р техн. наук, профессор, профессор кафедры программных средств Запорожского национального технического университета, Запорожье, Украина

МОДЕЛЬ И ИНДИВИДУАЛЬНЫЕ МЕТРИКИ КАЧЕСТВА НАУЧНЫХ ПУБЛИКАЦИЙ

Проанализированы известные метрики научных публикаций. Установлено, что их общим недостатком является оценивание статей на внешнем по отношению к их содержимому уровне, что не позволяет явно судить о качестве представления материала статьи. Целью данной работы являлась разработка комплекса показателей, позволяющих характеризовать свойства научных работ с точки зрения их структуры, лексики и формы представления материала, а также библиографии. Определен набор метрик, позволяющих количественно оценивать индивидуальные свойства материала и автоматизировать анализ научных публикаций. Предложенный набор включает показатели статьи: потенциал охвата читательской аудитории, структурированность, разнообразие географии, языков и видов источников библиографии, качество библиографии, насыщенность текста ссылками на источники, число и цитируемость рисунков и таблиц, иллюстрированность, объем и использование математического аппарата, эффективность аббревиатур, показатели лексики статьи (соответствие названия тексту, названия и авторской аннотации, авторской аннотации и текста, ключевых слов и аннотации, ключевых слов и названия, ключевых слов и текста, ключевых и характеристических слов, аннотации и характеристических слов, названия и характеристических слов, УДК и лексики статьи, описания таблиц и рисунков тексту, соответствия лексики абзацев, названий литературных источников и текста), самоцитирование авторами статьи, качество авторского коллектива, гибридные и интегральные показатели качества статьи. Приведены примеры, подтверждающие практическую применимость предложенных показателей.

Ключевые слова: наукометрия, библиометрия, качество, статья, научная работа, метрика, анализ цитируемости, важность статьи.

Subbotin S. A.
Dr.Sc., Professor, Professor of department of software tools, Zaporizhzhya National Technical University, Zaporizhzhya, Ukraine

MODEL AND INDIVIDUAL QUALITY METRICS OF SCIENTIFIC PUBLICATIONS

The existent metrics of scientific publications has been analyzed. Their common disadvantage is the external evaluation of articles relatively to external level to their content. This obviously does not allow to assess the quality of presentation of the article. The purpose of this paper is to develop a set of indicators to characterize the properties of scientific papers in terms of their structure, vocabulary and forms of presentation, as well as a bibliography. The set of metrics allowing to quantificate the material individual properties and to automate the analysis of scientific publications has been defined. The proposed set includes such paper indices as: potential readership coverage, structuring, diversity of geography, languages and types of bibliographic sources, bibliography quality, text richness by the links to sources, figures and tables number and quotability, Illustrativity, mathematical apparatus volume and the use, abbreviation effectiveness, paper vocabulary indices (compliance of title with text, of title with author's abstracts, of author's abstracts with text, of keywords with abstracts, of keywords with title, of keywords with text, of keywords with characteristic words, of abstracts with characteristic words, of names with characteristic words, of UDC with paper vocabulary, of tables and figures descriptions with text, of vocabulary compliance with paragraphs, of references names with text), self-citations of the authors, author team quality, hybrid and integrated indicators of paper quality. The examples confirming practical applicability of the proposed indicators are shown.

Keywords: scientometrics, bibliometrics, quality, paper, scientific work, metric, citation analysis, article importance.

REFERENCES

1. Moed H. F. Measuring contextual citation impact of scientific journals, *Journal of Informetrics*, 2010, Vol. 4, Issue 3, pp. 265–277. DOI: 10.1016/j.joi.2010.01.002
2. González-Pereira B., Guerrero-Bote V. P., Moya-Anegyn F. A new approach to the metric of journals' scientific prestige: The SJR indicator, *Journal of informetrics*, 2010, Vol. 4, Issue 3, pp. 379–391. DOI: 10.1016/j.joi.2010.03.002
3. Guerrero-Bote V. P., Moya-Anegyn F. A further step forward in measuring journals' scientific prestige: The SJR2 indicator, *Journal of Informetrics*, 2012, Vol. 6, Issue 4, pp. 674–688. DOI: 10.1016/j.joi.2012.07.001
4. Butler D. Free journal-ranking tool enters citation market, *Nature,* 2008, Vol. 451, Issue 6, pp. 6. DOI:10.1038/451006a.
5. Falagas M. E., Kouranos V. D., Arencibia-Jorge R., Karageorgopoulos D. E. Comparison of SCImago journal rank indicator with journal impact factor, *The FASEB Journal*, 2008, Vol. 22, Isuue 8, pp. 2623–2628. DOI:10.1096/fj.08-107938.
6. Bergstrom C. T., West J. D., Wiseman M. A. The Eigenfactor Metrics, *Journal of Neuroscience*, 2008, Vol. 28, Issue 45, pp. 11433–11434.
7. Hirsch, J. E. An index to quantify an individual's scientific research output. – Proceedings of The National Academy of Sciences, 2005, Vol. 102, No. 46, pp. 16569–16572. DOI:10.1073/pnas.0507655102.
8. Egghe L. Theory and practise of the g-index, *Scientometrics*, 2006, Vol. 69, No. 1, pp. 131–152. DOI:10.1007/s11192-006-0144-7.
9. Woeginger G. J. An axiomatic analysis of Egghe's g-index, *Journal of Informetrics*, 2008, Vol. 2, Issue 4, pp. 364–368. DOI:10.1016/j.joi.2008.05.002
10. Tol R. S. J. A rational, successive g-index applied to economics departments in Ireland, *Journal of Informetrics,* 2008, Vol. 2, Issue 2, pp. 149–155. DOI:10.1016/j.joi.2008.01.001.
11. Kosmulski M. I – a bibliometric index, *Forum Akademickie,* 2006, Vol. 11, P. 31.
12. Prathap G. Hirsch-type indices for ranking institutions' scientific research output, *Current Science*, 2006, Vol 91, Issue 11, P. 1439.
13. Igra v cyfir', ili kak teper' ocenivayut trud uchenogo (cbornik statej o bibliometrike). Moscow, MCNMO, 2011, 72 p.
14. Cyganov A. V. Kratkoe opisanie naukometricheskix pokazatelej, osnovannyx na citiruemosti, *Upravlenie bol'shimi sistemami : cb. tr. Spec. vyp. 44 – Naukometriya i e'kspertiza v upravlenii naukoj,* [pod red. D. A. Novikova, A. I. Orlova, P. Yu. Chebotareva]. Moscow, IPU RAN, 2013, pp. 248–261.
15. Shtovba S. D., Shtovba E. V. Obzor naukometricheskix pokazatelej dlya ocenki publikacionnoj deyatel'nosti uchenogo, *Upravlenie bol'shimi sistemami : cb. tr. Spec. vyp. 44 – Naukometriya i e'kspertiza v upravlenii naukoj*, [pod red. D. A. Novikova, A. I. Orlova, P. Yu. Chebotareva]. Moscow, IPU RAN, 2013, pp. 262–278.
16. List of languages by total number of speakers [Electronic resource]. Access mode: https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers.
17. Rekomendacii EASE (Evropejskoj associacii nauchnyx redaktorov) dlya avtorov i perevodchikov nauchnyx statej, kotorye dolzhny byt' opublikovany na anglijskom yazyke [E'lektronnyj resurs], 16 p. Rezhim dostupa: http://www.ease.org.uk/sites/default/files/ease_guidelines-june2014-russian.pdf
18. The UDC Summary [Electronic resource]. Access mode: http://www.udcc.org/udcsummary/php/index.php?lang=ru&pr=Y